University of Reading

Back to basics: how measures of lexical diversity can help discriminate between CEFR levels

Article

Accepted Version

Treffers-Daller, J., Parslow, P. and Williams, S. (2018) Back to basics: how measures of lexical diversity can help discriminate between CEFR levels. Applied Linguistics, 39 (3). pp. 302-327. ISSN 1477-450X doi: https://doi.org/10.1093/applin/amw009 Available at https://centaur.reading.ac.uk/54410/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1093/applin/amw009

Publisher: Oxford University Press

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

University of Reading

Central Archive at the University of Reading

Reading's research outputs online

Back to basics: how measures of lexical diversity can help discriminate between CEFR levels[1]

Jeanine Treffers-Daller*, Patrick Parslow** and Shirley Williams** (University of Reading)

*Institute of Education, ** Systems Engineering

Accepted for publication in Applied Linguistics

Back to basics: how measures of lexical diversity can help discriminate between CEFR levels[2]

Abstract

This study contributes to ongoing discussions on how measures of lexical diversity (LD) can help discriminate between essays from second language learners of English, whose work has been assessed as belonging to levels B1 to C2 of the Common European Framework of

[1] We are very grateful to Pearson Education Ltd for sponsoring us with a research grant that made this study possible, and to Kirsten Ackermann, Veronica Benigno and Jeremy Hancock for their support in carrying out this project.

[2] This work was funded by Language Testing, a trading division of Pearson Education Ltd. We are very grateful to Kirsten Ackermann, Veronica Benigno and Jeremy Hancock for their support in carrying out this project. We are also indebted to Scott Jarvis for providing a spreadsheet to help us with computing HD-D.

Reference (CEFR). The focus is in particular on how different operationalisations of what constitutes a "different word" (type) impact on the LD measures themselves and on their ability to discriminate between CEFR levels. The results show that basic measures of LD, such as the number of different words, the TTR (Templin 1957) and the Index of Guiraud (Guiraud 1954) explain more variance in the CEFR levels than sophisticated measures, such as D (Malvern et al. 2004), HD-D (McCarthy and Jarvis 2007) and MTLD (McCarthy 2005) provided text length is kept constant across texts. A simple count of different words (defined as lemma's and not as word families) was the best predictor of CEFR levels and explained 22 percent of the variance in overall scores on the Pearson Test of English Academic in essays written by 176 test takers.

Keywords: lexical diversity, CEFR, lemmatization, language testing, derivational morphology

1. Introduction

The words writers choose matter because the quality of a piece of writing depends to a large extent on the vocabulary that is deployed in it (Staehr 2008, Crossley and Macnamara 2011, Grobe 1981, Olinghouse and Wilson 2013). For this reason, researchers interested in the assessment of writing are looking for ways in which the vocabulary used in texts can be evaluated in objective and efficient ways. This is very important in the field of second language writing because there are so many second language learners, in particular of English, whose written work needs to be assessed. Automated measures which can give researchers, teachers and test developers a valid assessment of the vocabulary in students' work are particularly important for those interested in discriminating between texts of learners of different levels of language proficiency, as measured on the Common European

Framework of Reference (CEFR)[3]. Many researchers are trying to find criterial features whose presence or absence can differentiate between these levels albeit in the understanding that there is a great deal of overlap between levels as well as variability within individuals (Bartning, Martin, and Vedder 2010). Researchers working with measures of lexical diversity (LD), that is measures of the range of words deployed in oral or written language, can potentially make an important contribution to this field of research. This is because lexical diversity measures are often used as a general purpose measure of spoken and written language (Malvern et al. 2004) or as a measure of complexity at the lexical level (Housen, Kuiken, and Vedder 2012).

In this study we look at how different operationalizations of what constitutes the basic unit of measurement – the word, the lemma or the word family – affect the LD measures themselves and the ability of the measures to predict CEFR scores. We will compare a number of traditional measures with a number of newer, sophisticated measures hoping to find out whether the latter are better able to discriminate between different levels of achievement on the CEFR than the former, as any new measure should be informative over and above any existing measure if it is to have any incremental validity (McCarthy and Jarvis 2010, Treffers-Daller 2013). However, before we can make quantitative analyses of the words displayed in texts, it is important to look in detail at what constitutes a distinct word (a type). As Thomson and Thompson (1915, 54) put it, any statements about the number of

[3] The notion of CEFR level itself is problematic in that a particular learner is not necessarily at the same CEFR level with respect to vocabulary range, grammatical accuracy or phonological control. In addition, there is little empirical evidence for the links that are claimed to exist between CEFR levels and a range of existing standardised exams. This issue cannot be pursued here but the reader is referred to Alderson (2007), Hulstijn (2007) and Hulstijn, Schoonen, De Jong, Steinel, and Florijn (2012) for in-depth discussion.

words speakers or writers know "have no meaning unless a definite understanding exists as to what are and what are not different words". We need to know, for example, whether different inflected forms (*reads, reading, read*) or derived forms (*reader, readability, unreadable,* etc.) are counted as different tokens of one type or as different types. Under the first approach the unit of analysis is the lemma and under the second approach, it is the word family.

To the best of our knowledge no studies have been done into how lemmatization (grouping together different forms of a word by erasing affixes) affects the LD measures' ability to discriminate between learners of different levels on the CEFR. We hope the current article can help to clarify this issue.

The paper is structured as follows: In section 2 we sketch the discussion around the construct and the measurement of LD. In section 3 we look at various ways to define types (different words) in studies of LD. Section 4 presents the aims and research questions (4.1); these are followed by the methods (4.2) and the results (4.3) of the current project, and a summary (4.4). In section 4.5 we discuss the results of our analysis in the light of previous research and section 5 offers a final conclusion and outlook towards the future.

2. Lexical diversity: the construct and its measurement

Before we can start measuring LD in texts a few words must be said about the meaning of the construct. As Laufer and Nation (1995) and Durán et al. (2004) point out, LD is not just about the range of words a reader possesses, but also about the ways in which these words are deployed in texts. In a seminal paper on the nature of the construct of diversity, Jarvis (2013) develops this further and postulates that LD is a multidimensional construct which refers to readers' subjective perception of the LD of a text. The author argues that six different characteristics of a text contribute to a reader's perception of its lexical diversity, namely variability, volume, evenness, rarity, dispersion, and disparity (see Jarvis 2013 for details).

This is very important, because it means that analyses of the range of words deployed in a text only tap into one component of diversity, and certainly do not provide a comprehensive perspective on the lexical characteristics of a text. Lexical variability is, however, an important component of LD, as measures of lexical variability have frequently been found to correlate strongly with other text-internal measures as well as with text-external measures (see below for a discussion). This makes it even more important to take lexical variability seriously and to explore different ways to measure it in more depth. In this study we will continue to use the term lexical diversity (LD) as this is the term most widely used for the range of words deployed in a text, but are aware that we are in fact only measuring one aspect of lexical diversity, namely lexical variability.

The best known of the traditional measures of LD is the Type-Token Ratio (TTR), which expresses the ratio of different words (types) to total words (tokens) in a given language sample. This measure is often attributed to Templin (1957) but is already mentioned by (Johnson 1944). Thus, if a text consists of 100 tokens, and 72 types, the TTR is 72/100 - .72. The key problem with the TTR is, however, that "TTR's for samples of different magnitudes are not directly comparable because of the tendency for the TTR to vary inversely with the size of sample" (Johnson 1944, 2). In other words, the longer the sample, the lower the TTR. More recently other measures of LD were developed, such as D (Malvern et al. 2004), HD-D (McCarthy and Jarvis 2007) and MTLD (McCarthy 2005). The D-value is computed through a series of computations of the Type-Token Ratio (TTR) on samples of different text lengths (typically ranging from 35-50 tokens) after which a random sampling TTR curve is computed (but see McCarthy & Jarvis, 2007, for a critical appraisal of D). HD-D is similar to D but based on the hypergeometric distribution function (Wu 1993). HD-D calculates, for each lexical type in a text, the probability of encountering any of its tokens in a random sample of 42 words drawn from the text (McCarthy and Jarvis 2010, 383).

The measure of textual diversity (MTLD) was developed by McCarthy (2005) and later tested by Crossley, Salsbury, and Macnamara (2009), McCarthy and Jarvis (2010) and Treffers-Daller (2013). This measure is calculated as the mean length of sequential word strings in a text that maintain a given TTR value, which McCarthy and Jarvis (2010) have chosen to be 0.720. MTLD calculates the TTRs in a sentence until the TTR drops to 0.72, at which point the first factor is complete and TTRs are counted from scratch again: as in the following example: *of* (1.00) *the* (1.00) *people* (1.00) *by* (1.00) *the* (.800) *people* (.667) |||FACTORS = FACTORS = 1||| *for* (1.00) *the* (1.00) *people* (1.00) . . . and so on (McCarthy and Jarvis 2010, 384).[4] Subsequently MTLD is obtained by dividing the total number of words by the total number of factors. Thus, if the text is 360 words long and there are 4 factors, the MTLD value is 90.

As shown in Author (2013) text length dependency is also a problem for the newer measures of LD. Whilst D and HD-D increase with text length, MTLD decreases with text length. The search for a measure of LD which is not dependent on text length is still on, a century after the first studies in this field appeared (Thomson and Thompson 1915), but the search for a measure that can be used with texts of any length, for spoken and written data, and any register (every day speech as well as academic language), has so far proved elusive and is somewhat similar to the search for the Holy Grail (Malvern et al. 2004). In this context, it is important to note that despite the tremendous efforts employed in the development of new measurements, very simple counts of the number of types can sometimes be more successful than complex formulae, such as D, in detecting differences between group or within group differences. Tonkyn (2012) demonstrates that neither subjective ratings of lexical complexity, nor objective measurements of LD were able to

[4] The example is kept brief for reasons of space. Factors do not normally consist of so few words.

detect gains in lexical complexity in students' speaking skills. The only measures that were significantly different were measures of lexical sophistication: simple counts of the number of rare types and the number of rare families. A similar point has been made by Richards and Chambers (1996). That the number of types is a good measure of LD also emerges from the study of Lancashire and Hirst (2009), who used this measure to analyse the longitudinal development of LD in fourteen novels written by Agatha Christie between the ages of 28 and 82. They kept the number of words analyzed constant at 50,000 for each novel and showed that LD decreases significantly from the early to the late novels. Lancashire and Hirst argue that analyses of LD over time in the writing of patients may be a valid tool for the diagnosis of different forms of dementia.

While research into the internal structure of the construct of LD is only just starting, there is considerable evidence in the available literature regarding the relationships between LD and a range of text-internal or text-external variables. Summarizing the discussion about the measurement of LD is not easy because researchers use different measures and different statistical tests (e.g. regression analysis, discriminant function analysis or simple correlations) to investigate the relationships between variables. To begin with *text-internal* measures, Grobe (1981) is probably one of the first studies which systematically investigated the relationship between a range of automated text characteristics, including LD measures, and holistic ratings of essay quality. She found that measures of LD were among the top nine predictors of holistic ratings of essay quality. At grade eleven, it was the number of types which explained 29.3% of holistic essay ratings. Yu (2010) also found correlations of .33 between D (Malvern et al. 2004) and a holistic assessment of students' writing performance, and an even stronger correlation of .48 between D and holistic assessments of their oral performance.

Different reasons have been advanced for the relationships between holistic ratings and measurements of vocabulary diversity of texts. Of course it is likely that, as Laufer and Nation (1995, 307) note, "a well-used rich vocabulary is likely to have a positive effect on the reader", but this does not explain how such positive impressions arise in readers, nor how raters use this information. Tonkyn (2012) advances an interesting hypothesis, suggesting that these impressions might be the result of raters consciously or subconsciously counting the number of different words used in an essay or an oral performance, although this remains difficult to prove with currently available research tools.

Some of the researchers working in this area use LD measures to get a better understanding of the complexity of texts (or spoken language). These studies are part of the field of complexity, accuracy and fluency (CAF) in the production of learners. Bulté and Housen (2012) make a major contribution to this field by providing a model of the different constructs under study (see Figure 1). In their model LD is situated at the middle level as one of the three behavioural constructs which contributes to what they call systemic lexical complexity (at the top-level of theoretical constructs). This model is particularly helpful to develop our understanding of the construct of LD because it makes a link between LD and complexity by considering LD as a specific component of lexical complexity.

Figure 1 approximately here

Researchers working on correlations between LD measures and *text-external* variables are generally more interested in the properties of learners than in the properties of texts. They want to know how LD measures correlate with or can predict independent measures of general language proficiency (or competence), such as scores on a receptive vocabulary test (Vermeer 2000). Daller, Van Hout and Treffers-Daller (2003) found

correlations ranging from .40 and .45 between LD scores and C-test scores for German as found among Turkish-German bilinguals. The existence of even stronger correlations (from .56 to .79) between LD and a French C-test among British learners of French were found in Treffers-Daller (2013). Further evidence regarding the link between LD and overall language proficiency comes from Crossley, Salsbury, and Macnamara (2013) who studied essays written by three groups of L2 learners which were created on the basis of TOEFL scores. They reported that a measure of LD (Maas 1972) was the second strongest predictor of overall language proficiency (after word imagability, that is the extent to which a word evokes mental and sensory images, Crossley and Macnamara 2011) in a Discriminant Function Analysis. In a follow-up study, Crossley, Salsbury, and Macnamara (2014) report correlations in the range of .8 and .9 between holistic ratings of essay and discrete ratings of LD.[5] Thus, there is considerable evidence that LD measures can be used as a proxy for general language ability. If this is indeed the case, measures of LD should be able to predict overall CEFR scores in that they explain a significant proportion of CEFR scores for a group of L2 learners.

3. What is a type? The issue of lemmatization

Researchers in the field have taken very different approaches to the issue of what constitutes a different word (type). A large number of researchers, including Engber (1995), Laufer (1991) and Jarvis (2002), consider different inflected forms as tokens of the same type. Durán et al. (2004) opt for the same approach but consider fused forms (such as *fell* – past tense of *fall*) as separate types. Other researchers use a different method. Yu (2010), for example,

[5] As the correlations between ratings are very strong, one wonders whether raters were really giving independent ratings of the different constructs or whether a halo effect is the true cause of the strength of these correlations, but this issue cannot be pursued here.

considers all inflected forms as different types, and the same approach is taken by the Vocabprofile programme (http://www.lextutor.ca/vp/comp/) on the Lextutor website developed by Tom Cobb on the basis of Nation's Range programme (http://www.victoria.ac.nz/lals/about/staff/paul-nation). Laufer and Nation (1995), on the other hand, use the word family as the unit of analysis and include all inflected forms as well as derivational forms listed in Nation and Bauer (1993), up to level 3 (*-able, -er, -ish, -less, -ly, -ness, -th, -y, non-* and *un-*). Thus, forms such as *governable* and *ungovernable* are considered as tokens of the type *govern*, but *government* is a separate type, because the suffix *-ment* belongs to a higher level of the scale. Finally, there is a large group of researchers who do not explain whether they have lemmatized the data. If the research in this field is to make any progress, and particularly if we are to clarify which LD scores can be expected at different levels of the CEFR, we need to be clearer about what we count and how we count. As Durán et al. (2004) have shown, different lemmatization principles result in different LD scores, and there are significant differences between LD scores for text that have been lemmatized and those that have not been lemmatized (Treffers-Daller 2013).

Evidence from language processing studies can throw new light on the issue of lemmatization too. For English it has been shown that L2 learners do not automatically have productive knowledge of all the derived forms if they master the stem form (Schmitt and Zimmerman 2002). Thus, if an L2 learner has problems producing the derived form *selection* on the basis of the stem *select*, these two are apparently separate words for learners, which could indicate it is probably *not* a good idea to consider the derived forms of a word as one type in analyses of LD in L2 learners. Other psycholinguistic evidence points into the same direction. Morphologically complex forms such as frequent and regular past tense forms in *–ed*, for example, are not always processed through a decompositional process that segments a word into its different morphemes, but instead may be retrieved through direct look-up of a

whole-word representation stored in lexical memory (Silva and Clahsen 2008). There are also interesting differences between L1 and L2 users in the processing of morphologically complex forms. In a priming study, Silva and Clahsen (2008) demonstrated, for example, that respondents are quicker identifying that *bitter* is a word if they have been primed by a derived form such as *bitterness*, but the stem priming effect is much less strong in L2 than in L1. The authors conclude that L2 processing relies less on morphological decomposition than L1 processing. For the current study, the psycholinguistic evidence discussed here probably means that we cannot take it for granted that L2 learners know there is a link between a root and a derived or an inflected form, and some inflected and derived forms may be stored as unanalyzed wholes in memory. The implications for studies of LD are that it may be not be appropriate to apply extensive lemmatization.

Whether or not lemmatization is appropriate also depends on the aims of the research. If researchers want to be able to compare lexical ability across typologically different languages, as in Treffers-Daller and Korybski (2016) then lemmatization is essential, because it helps to make LD scores in different languages more comparable. However, if the aim is to predict CEFR levels on the basis of LD measures then lemmatization may be counterproductive because important information regarding a learner's ability to form complex words and to use inflections is lost in the process of lemmatization.

4. The current study

4.1 Aims and research questions

In this study we therefore look at how different operationalizations of what constitutes the basic unit of measurement – the word, the lemma or the word family – affect the LD measures themselves and the ability of the measures to predict CEFR scores. In addition we

wanted to find out to what extent text length (number of tokens) affected the scores test takers obtained for their essays. The current study focuses in particular on the following questions:

1) What is the effect of different types of lemmatization on the LD scores?
2) How do different lemmatization principles affect the ability of the LD measures to discriminate between CEFR levels?
3) Can newer measures of LD explain additional variance in CEFR scores over and above traditional measures (incremental validity)?
4) To what extent does the length of the students' essays contribute to the overall scores for the essays as given by Pearson?

4.2. Methodology

4.2.1 Participants

The participants in the study were 179 adult learners of English from 47 different countries. The largest number originated from India (39), Nigeria (17), Pakistan (19), and the Philippines (11). The total number of different languages reportedly spoken at home was 39. Among these English was spoken most frequently (67), followed by Urdu (15), Tagalog (10), Hindi (9), French (7) and Spanish (6), with other languages spoken by 5 participants or less. The participants' ages ranged from 16-51 and their mean age was 26.82 (SD 6.40). 68 of the participants were female (38%) and 111 were male (62%).

All students had written an essay as part of taking the Pearson Test of English Academic (PTE Academic), which is a computer-based academic English language test for international study, calibrated to the levels of the CEFR (from B1 to C2). Students had been allocated to a

particular CEFR level on the basis of a wide range of assessments, which involved twenty different item types. From Pearson's database, 50 students were selected whose work was classified as B1, B2 or C1 on the basis of the overall test results. For the C2 level, the number of students included in the study was 29 as there were too few students who had reached this level in the database (see Table 1).

Table 1 approximately here

4.2.2 Materials

The data consisted of essays written by students on one of two different topics, for which two prompts were provided, which cannot be disclosed for confidentiality reasons. For levels B1 to C1 25 students wrote an essay based on topic 1 and 25 students on topic 2. For level C2 these figures were 15 and 14 respectively. Students had to write between 200 and 300 words for this essay within 20 minutes. The mean length of students' essays was 249.47 words (SD 35.70), with the minimum being 187 and the maximum 357. As is explained in the PTE Academic Score Guide (http://pearsonpte.com/wp-content/uploads/2014/07/PTEA_Score_Guide.pdf), each test taker completes between 70 and 91 items in any given test and there are 20 different item types. Total scores range from 10–90 points. Pearson provided us not only with the essays but also with students' scores on the PTE Academic. These were the test takers' total scores, as well as their scores on a number of discrete variables: their vocabulary score (based on 15 items), their writing score (based on 15 items) and their item score for the essay.[6] The latter was based on a range of different

[6] The scores provided by Pearson are based on a range of tasks, which include the item score for the essay on which the LD scores were computed. It would have been preferable to use scores which were completely independent of the texts on which the LD scores were

criteria, including content, spelling, grammar, vocabulary etc. Further details about the contribution of different variables to the different scores and the test procedure can be found on the PTE Academic website, http://pearsonpte.com/PTEAcademic/Pages/home.aspx). We decided to keep text length constant, as text length affects measures of LD (see section 1). A cut-off point of 200 tokens was chosen to include as many essays as possible whilst maintaining a sufficiently long text, because calculating LD on very short texts is problematic (Jarvis 2002, Treffers-Daller 2013). This way only three essays from the 179 were discarded (one from the B1 and two from the C1 bands). We used McCarthy's gramulator to select 200 words from the middle of each essay to maximize opportunities for inclusion of parts of the beginning, the middle and end of the essays in the chosen extract, whilst avoiding sentences with copies of the prompt, which is a strategy in particular of lower level learners (Kobrin, Deng, and Shaw 2011). Such copies of lexical and syntactic structures from the prompt were found to occur most frequently at the start of essays.

4.2.3 Data treatment

Prior to data analysis, we changed the format of the data to CHAT (MacWhinney, 2000) and marked all proper names, acronyms, cardinal numbers[7] and incomprehensible words with an

---

calculated, but Pearson could not provide these. However, this is not pertinent for the current research questions, which involve a comparison of the ability of different measures of LD to predict scores, as all LD scores are affected in exactly the same way.

[7] Cardinal numbers were excluded because knowledge of these numbers does not reflect language knowledge, contrary to knowledge of ordinal numbers in *–th*, which were included in the analysis (see also section 4.2.4).

ampersand (&) to indicate these should be ignored in further analyses by CLAN (MacWhinney 2000). These were excluded because if speakers use these frequently this leads to an inflation in LD scores and there is a risk that the latter are no longer a valid index of a student's language ability. We also corrected spelling mistakes in the data, using the CHAT conventions. It is important to clean the data prior to calculating LD, because learners often make spelling mistakes; if the same word occurs in a text once with a spelling mistake (e.g. *goverment* instead of *government*) and once in the correct spelling, programs such as CLAN or any other software will count these as different words, which is clearly undesirable. Abbreviations such as TV were written out in full as *television*. In what follows the numbers in parentheses refer to the numbers we gave to transcripts in CLAN. We also excluded non-existing words, such as *trustful* or *solutionated* without replacing these with existing words. There were only 20 mistakes of this kind in the data set. We did not correct the wrong uses of existing words, such as *in* where *on* would be expected or the omission or incorrect use of determiners.

4.2.4 Lemmatization and data analysis

There is some evidence Treffers-Daller (2013) that measures of LD are better at discriminating between L2 of French of different levels if the data are lemmatized, but as explained in section 3, there are also disadvantages to lemmatization. In this study we tried out two different ways to lemmatize the data. First of all, we took the LEMMA as the unit of analysis. If this first lemmatization principle is used, all inflected forms of verbs, nouns and adjectives are considered to be tokens of the same type. Thus, for example, *work, working, works* and *worked* are considered to be tokens of the same type, but *worker* and *workable* are different types, because they are constitute a different lemma: they are derived from the root *work* through the addition of derivational suffixes. Second, the WORD FAMILY was chosen as

the unit of analysis. Under the second principle, all inflected forms and the derived forms up to level 3 in Nation and Bauer (1993) are considered to be tokens of the same type.

Prior to the analysis all data were converted to CHAT format and tagged with morphosyntactic codes on a separate tier. Subsequently the data were analysed with the CLAN tools. Example (1) illustrates the ways in which CLAN tags the data on the morphosyntactic tier (mor tier). The main tier is the line with the asterisk (*) followed by the three letter code XXX which represents the speaker/writer. The line marked %mor is the mor tier.

(1)

*XXX: John worked on one project until yesterday and wanted to work on
the other project today but to his utter surprise his fellow
workers had already finished it .
%mor: n:prop|John v|work-PAST prep|on det:num|one
n|project prep|until adv:tem|yesterday coord|and v|want-PAST inf|to
v|work prep|on det|the qn|other n|project adv:tem|today conj|but
prep|to pro:poss:det|his adj|utter adj|surprise pro:poss:det|his
adj|fellow n|work&dv-AGT-PL v|have&PAST adv|already part|finish-PASTP
pro|it .

If no lemmatization is applied (from now on lemma 0), different forms of the same verb (e.g. *worked* and *work* in (1)) as well as derivations such as *worker* will be considered as different words (types). This means that if types and tokens are counted according to the lemma 0 principle, example (1) consists of 28 tokens (total number of words), and 24 types (number of different words), because *his, on, project* and *to* are repeated two times each. Under the

first lemmatization principle (from now on lemma 1), *worked* and *work* will be considered tokens of the same type, but *worker* will be seen as a different type. Thus, according to the lemma 1 principle, there are 28 tokens but 23 types (*his, on, project, to* and *work* occur twice each). Finally, under the second lemmatization principle (from now on lemma 2), *worker* will no longer be considered as a separate type. Therefore if the second lemmatization principle is adopted there are 28 tokens but only 22 types in example (1).

The majority of the affixes which are coded on the mor tier belong to level 2 or level 3 in Nation and Bauer (1993). This is important because it means that only affixes from these two levels will be erased in the process of calculating LD measures on lemmatized files. The level 2 affixes are the inflections (tense and aspect on verbs and number on nouns) whilst level 3 comprises the most frequent and regular derivational affixes: *-able, -er, -ish, -less, -ly, -ness, -th, -y, non-, un-*. As Bauer and Nation point out, the existence of homographs complicates the classification. Fortunately the coding on the mor tier also makes a distinction between some of these homographs. One of the two different *–ly* affixes, for example, forms adverbs on the basis of adjectives, such as *sadly*, which is coded as adv|sad&dadj-LY, and the other one forms adjectives on the basis of nouns, as in *leisurely*, which is coded as adj|leisure&dn-LY. Only the former belongs to level 3 in Bauer and Nation's list. There are also two different *–th* suffixes: the *-th* which turns cardinal numbers into ordinal ones, as in *tenth*, belongs to level 3, whereas nouns such as *growth* or *strength*, which are derived from verbs or adjectives, are listed under level 6. As *-th* is not coded separately on the mor tier the distinction between the two *–th* suffixes could not be made in our analysis. However, there were only three ordinal numbers ending in *–th* in the data set (*billionth*, *fourth* and *twentieth*). In only one case (*billionth*) did the cardinal number occur in the same text as the corresponding ordinal number (*billion*). The ordinal and the cardinal numbers were not counted as separate types (which would have been inappropriate) because all cardinal

numbers were excluded from the analysis and only ordinal numbers were included (see 4.2.3). There were 37 nouns ending in *–th* which were derived from verbs or adjectives in the data. As the mor tier in CHAT does not mark the suffix *–th* on the dependent tier, these suffixes were not erased in the process of calculating the measures of LD. All 37 nouns were therefore counted as separate types for the purpose of LD analyses. This was appropriate because these derivations belong to level 6 in Bauer and Nation (1993) and the derivational suffixes of these words should therefore not be erased.

It is also important to note that it makes a difference whether types are counted on the main tier or on the mor tier. For the purposes of the current study, it is interesting to carry out analyses on the mor tier because this tier can be used to distinguish between homographs such as prepositional *to* as in *to his surprise* and infinitival *to* as in *wanted to work* in (1). If lexical richness analyses are carried out on the mor tier in a CHAT transcript, CLAN will consider these two uses of *to* as different types because different codes are allocated to each (prep|to versus inf|to). Thus, if the first lemmatization principle is applied and analyses are carried out on the mor tier, there are more types in a text than when the analysis is carried out on the main tier because on the main tier no distinction is made between homographs and CLAN will analyse both forms of *to* as tokens of the same type, whereas on the dependent tier these will be considered as different types.

We used the following command in CLAN to compute VOCD following the first lemmatization principle:

vocd +t%mor -t* +s"*|*-%%" +s"*|*&%%" +s"*|*~%%*|*" +d3 @[8]

[8] This command tells CLAN to compute VOCD on the mor tier after erasing any codes representing inflectional or derivational affixes which are marked with hyphens, ampersands or tildes. The d3 switch sends the output to an excel spreadsheet.

Example 2 shows what the output from VOCD looks like after the erasure of all derivational and inflectional affixes:

(2)

n:prop|john v|work prep|on det:num|one n|project prep|until adv:tem|yesterday coord|and v|want inf|to v|work prep|on det|the qn|other n|project adv:tem|today conj|but prep|to pro:poss:det|his adj|utter adj|surprise pro:poss:det|his adj|fellow n|work v|have adv|already v|finish pro|it

Note that the word classes in (2) are maintained after the erasure of affixes. This way, *work* and *worker* are considered as separate types because *work* is marked as a verb and *worker* as a noun (despite the fact that the *–er* suffix has been erased).

For the calculation of VOCD with the second lemmatization principle, where the unit of analysis is the word family, we used the following command:

vocd +t%mor -t* +s"@r-*,o-%" @ +d3

The use of this command leads to erasure of all information regarding word categories and affixes. As a result, *work* and *worker* are no longer distinguished and seen as tokens of the same type, and homographs such as infinitival *to* and prepositional *to* are no longer distinguished either.

The other indices of LD were computed as follows: the Index of Guiraud (types/ √ tokens) was calculated with SPSS on the basis of the counts of types and tokens provided by CLAN under the different lemmatization principles described above. As HD-D and MTLD

are not available under CLAN, we used a spreadsheet provided by Jarvis to compute HD-D and McCarthy's gramulator (https://umdrive.memphis.edu/pmmccrth/public/software_index.htm) to compute scores for MTLD. Because the gramulator cannot handle the CLAN codes in the data (pers. communication Phil McCarthy, 12thApril 2013), we created three different lemmatized versions of all texts: a non-lemmatized version (lemma 0), and two versions which were lemmatized according to the two principles mentioned above (lemma 1 and lemma 2), with the help of two include files, named lemm1.cut and lemm2.cut. Lemm1.cut listed all inflected forms, and lemm2.cut all the inflected and derived forms up to level 3 in Bauer and Nation's list. With the change string command we then replaced all inflected/derived forms with root forms to create lemmatized files according to the two different lemmatization principles.

Because there could be slight differences[9] between the computation of, for example, D on the mor tier and on the main tier, we decided to compute D (and the number of types) on the main tier as well as on the dependent tier. Comparisons between indices of lexical diversity could thus be done on exactly the same versions of each file. For the analysis we used parametric tests because all variables (LD scores as well as the Pearson CEFR scores) were normally distributed, as revealed by a Kolmogorov-Smirnov test.

4.3. Results

4.3.1 The effect of lemmatization on the measures of lexical diversity

[9] In the version we created with the help of change string, differences between homographs (for example the differences between infinitival *to* and the preposition *to*) could not be made so the number of different types could be lower in this version than in the version computed on the mor tier.

The most basic measure of lexical diversity is a simple count of the number of different types in each text, and all measures of lexical diversity reported in this study use types as the basic unit of analysis. Therefore we will first provide information about the effect of the different lemmatization techniques on the number of types in the texts before answering our research questions.

When the number of different types is counted on the mor tier, the mean number of types per informant across all CEFR levels is 108.72 types (lemma 0). The corresponding figures for the two different lemmatized versions are 108.55 (lemma 1) and 101.56 (lemma 2). As could be expected, the number of types is higher for the unlemmatized version than for the lemmatized versions, but only the differences between lemma 0 and lemma 2 and between lemma 1 and lemma 2 are statistically significant (Repeated measures ANOVA, $F=839.40$, $df =2$, $p <.001$).[10] Type counts on Lemma 0 and lemma 1 are not significantly different from each other. However, if types are counted on the main tier, the number of types in the lemmatized versions is lower (because of the lack of distinction between homographs). For lemma 1 the mean number of types is then 103.43 and for lemma 2 it is 103.21 (see also Table 1). The differences between the lemmatized and the non-lemmatized versions are now all significant (Repeated measures ANOVA, $F=766.72$, $df = 1.134$), $p<.001$).[11] As computing the number of types on the main tier led to significantly different results between all three different lemmatization principles, whereas counting the types on the mor tier did not, counting types on the main tier was considered most interesting for a test of the effect of

[10] The Mauchly test for sphericity was not significant, so no correction for sphericity was needed. Post hoc tests were adjusted with the Bonferroni correction, and in all subsequent calculations so this will only be mentioned here once.

[11] The Mauchly test for sphericity was significant ($\chi^2=254.98$, $df =2$, $p <.001$). Therefore degrees of freedom were adjusted using Greenhouse-Geisser estimates of sphericity ($\varepsilon=.567$).

lemmatization on a) the LD measures themselves and b) the ability of these measures to predict differences in CEFR scores. We therefore used the counts from the main tier in all calculations. An added advantage of this approach is that all measures of LD are calculated on exactly the same version of the texts.

4.3.2 Lemmatization of LD measures and the CEFR levels

The results for the three different variants of the LD measures for each CEFR level have been split in Tables 2a (basic measures) and 2b (sophisticated measures) for ease of reference. Table 2a shows that mean scores for all basic measures are higher for the higher CEFR levels than for the lower ones, which is the expected as those at higher CEFR levels should have a richer vocabulary than those who obtained lower CEFR scores. With respect to the different lemmatization principles, Table 2a reveals that for TTR, the Index of Guiraud and D, calculations based on the lemma 0 version obtains the highest score, followed by lemma 1, while the lemma 2 version receives the lowest scores, as would be expected as second lemmatization principles reduces the number of types to a larger extent than the first lemmatization principle. For TTR, the overall differences between the scores for the three lemmatization principles are statistically significantly different (F= 1253.654, df 1.2, p <.001)[12] and the same is true for the Index of Guiraud (F =1064.2332, df = 1.172, p <.001)[13]. The similarity between the results for these three basic measures of LD is not surprising given the fact that they are mathematical transformations of each other.

[12] The Mauchly test for sphericity was significant (χ^2=191.70, df =2, p <.001). Therefore degrees of freedom were adjusted using Greenhouse-Geisser estimates of sphericity (ε=.602).

[13] The Mauchly test for sphericity was significant (χ^2=216.79, df = 2, p <.001). Therefore degrees of freedom were adjusted using Greenhouse-Geisser estimates of sphericity (ε=.586).

As for the sophisticated measures, given in Table 2b, the results are similar to those in Table 2a in that the scores for all sophisticated measures increase with CEFR level, but the different lemmatization principles do not present a coherent picture. For D, the lowest scores are generally obtained for the lemma 2 (except for the C1 level), whilst for HD-D and MTLD the lemma 2 version obtains the highest scores. The fact that the overall mean scores obtained for lemma 2 are *higher* than those for lemma 1 is counter intuitive. If there are fewer types in the text due to more rigorous lemmatization, one would expect these scores to be lower. The overall differences between the three lemmatization principles are significant for D (VOCD) in a repeated measures ANOVA, ($F=678.653$, $df = 1.11$, $p<.001$),[14] but the post hoc tests show that only the differences between lemma 0 and lemma 2 and between lemma 1 and lemma 2 are significant, but not the differences between lemma 0 and lemma 2. For HD-D all scores are significantly different from each other ($F=894.17$, $df = 1.65$, $p <.001$).[15] Finally, for MTLD all scores are again significantly different from each other ($F =232.653$, $df = 1.30$, $p <.001$).[16]

The Eta Squared values in Tables 2a and 2b clearly reveal that measures based on the first lemmatization principle explain most of the variance in scores at the different CEFR levels. The sophisticated measures have considerably lower Eta Squared values, but among this group of measures, the MTLD is the most powerful one, with an Eta Squared value of (.140). As the first lemmatization principle is the most successful one in explaining variance

[14] The Mauchly test for sphericity was significant ($\chi^2=294.64$, $df = 2$, $p <.001$). Therefore degrees of freedom were adjusted using Greenhouse-Geisser estimates of sphericity ($\varepsilon=.552$).

[15] The Mauchly test for sphericity was significant ($\chi^2=42.24$, $df = 2$, $p <.001$). Therefore degrees of freedom were adjusted using Huyn-Feldt estimates of sphericity ($\varepsilon=.832$)

[16] The Mauchly test for sphericity was significant ($\chi^2=138.14$, $df = 2$, $p <.001$). Therefore degrees of freedom were adjusted using Greenhouse-Geisser estimates of sphericity ($\varepsilon=.649$)

across different levels of the CEFR we have chosen this one to test whether the scores at different CEFR levels are significantly different from each other.

Table 2a approximately here

Table 2b approximately here

We then analysed the differences between the LD measures (based on the first lemmatization principle) across different levels of the CEFR (Table 3). Significant differences were found for all three basic measures for differences between B1 and B2, B1 and C1 and B1 and C2. Among the sophisticated measures only MTLD discriminated significantly between B1 and C1 and B1 and C2. The scores for D and HD-D only discriminated between the lowest (B1) and the highest CEFR levels (C2). The LD scores for B2 were not significantly different from the two highest levels (C1 and C2) for any of the measures, and the same was true for the differences between C1 and C2. The highest F-value was obtained by TTR and the lowest by D. Figure 1, which presents the different values obtained for Guiraud (lemma 1), also makes it clear that there is a considerable amount of overlap between scores at different levels.

Finally we investigated whether disambiguating between homographs made a difference for the ability of the measures to discriminate between CEFR levels. We found higher Eta squared values for LD measures which did not disambiguate between homographs, which means that making this extra effort was not helpful.

Table 3 approximately here

4.3.3. LD measures as predictors of CEFR scores

In addition to the overall CEFR level allocated to each student, Pearson provided us with a vocabulary score, a writing score and an overall score. Because these are interval data and the variables are normally distributed it is possible to compute Pearson correlations between these scores and the LD measures. In addition, this will provide information about the correlations between LD measures, which should be high, given the fact that they all measure the same construct, albeit in slightly different ways. Table 4 shows that this is indeed the case: the LD measures correlate strongly and significantly with each other. The basic measures correlate so strongly with each other (r >.97) that they can be considered to be identical. Among the sophisticated measures, the strongest correlation was found between HD-D and D (r = .925), which indicates that the measures are virtually identical too, although the metrics used to obtain these are different. There are also moderate correlations between the basic LD measures and the Pearson scores: out of all LD measures, the types correlate most strongly with the overall score (r =.470) and with the writing score (r =.447), whilst Guiraud correlates most strongly with the vocabulary score (r =.472). Among the sophisticated measures, MTLD correlates most strongly with the Pearson scores: writing score (r =.344), overall score (r =.338), and vocabulary score (r =.331).

Table 4 approximately here

We then carried out a regression analysis with the Pearson scores as dependent variables and two LD variables as predictors. As the basic variables correlate too strongly with each other to be entered together in a regression analysis, we chose one from the group of basic LD measures and one from the group of sophisticated measures, that is the ones which correlated most strongly with either the overall score, or the writing score or the vocabulary score. The collinearity statistics show that the tolerance and VIF are within the limits for allowing these

two variables to be entered together.[17] The results show that the number of types was the best predictor of the three Pearson scores. In a regression analysis with the overall score as the dependent variable, the number types predicted 22 percent of the variance in overall score (F (1, 174) = 49.293, p <.001; *b* = .470; R^2 =. 216). The number of types used also predicts the writing score (F (1, 174) = 43.450), p <.001; *b* = .447; R^2 =.195) and the vocabulary score (F (1, 174) = 48.783, p <.001; *b* = .468; R^2 =. 214). If MTLD is added to this model as an additional predictor, it turns out not to be significant. However, when entered on its own MTLD is a significant predictor of the overall score, but it explains less of the variance than the types (F (1, 174) =22.43, p <.001; *b* = .338; R^2 =.109). Similar values were obtained when MTLD was used as a predictor on its own in a regression analysis with the vocabulary score (R^2 =.104) or the writing score (R^2 =.110) as the dependent variable.

Finally, we were interested in finding out whether unabridged text length (that is the original total number of tokens for each essay) could predict the final score obtained by students. This was indeed the case (F(1, 174) = 39.850, p <.001, *b* = .432, R^2 = .182). We then entered both the types (lemma 1) and the tokens into a regression analysis. This was possible because the collinearity statistics were within the accepted limits.[18] We found that both made an independent and significant contribution to the variance in students' overall scores (F(2, 173) = 43.942, p <.001, *b* (tokens) = .349 and *b* (types) = .397). By comparison with using both predictors in separate regression analyses, the explained variance rises considerably to R^2 = .329. Thus, text length and lexical diversity taken together explained no less than a third of the variance in total score. Regression analyses which had the writing score or the vocabulary as the dependent variable, and the types and the tokens as predictors

[17] Tolerance: .387; VIF: 2.583

[18] Tolerance: .956; VIF: 1.046

produced The R^2 values of .314and .248 respectively. Again both the types and the tokens made independent contributions to the variance in the dependent variables.

Figure 2 approximately here

4.4 Summary of the results

This project has shown, first of all, that lemmatization has a significant effect on LD scores. We created three different versions of all texts: one with no lemmatization (lemma 0), one which was lemmatized according to the first principle (lemma 1), based on the lemma as the unit of analysis, and one which was lemmatized according to the second principle (lemma 2), based on the word family. All measures were then computed on 200 tokens selected from the middle of each transcript, so that text length was kept constant. For all LD measures investigated in this study, the scores are significantly lower if lemmatization is applied than if no lemmatization is applied.

As expected scores on the LD measures were generally higher for test takers whose overall performance on the CEFR was better. In addition, for most measures (except HD-D and MTLD) scores computed on the lemma 1 versions of the texts, were higher than those computed on the lemma 2 versions of the texts. Eta Squared values clearly showed that the first lemmatization principle explained most of the variance in scores at the different CEFR levels. Because the lemma-based lemmatization principle was found to be more useful to disambiguate student levels on the CEFR than the word family-based principle or using texts that were not lemmatized, we concluded that lemmatization should not erase derivational affixes, as they provide important information about test takers' knowledge. We also found that disambiguating between homographs did not add value to the measures' ability to

distinguish between CEFR levels because measures which did not take homographs into account obtained higher Eta Squared values than those that did.

Finally, the results showed that the Eta Squared values were higher for the basic measures (types, TTR and Guiraud) than for the sophisticated measures (D, HD-D and MTLD). Among the sophisticated measures, MTLD was best able to distinguish between CEFR levels. The basic measures correlated also more strongly with the Pearson overall score, the writing score and the vocabulary score than the sophisticated measures. The correlations among the basic LD measures themselves were so strong ($r > .97$) that they could arguably be considered to be virtually identical.

4.5 Discussion

The results summarized in the previous section underline the importance of the statement made over a century ago by Thomson and Thompson (1915, 54) that we need to be clear about what a different word is, if we want to have a meaningful discussion about the number of words used by speakers or writers. The current study has provided clear evidence that inflected forms (*works*, *working*, *worked*, etc.) should indeed be considered to be tokens of the type *work*, whilst derived forms such as *worker* or *workable* should *not* be considered to be tokens of the type *work*, but different types altogether, at least for the purpose of investigations involving measures of LD which focus on predicting CEFR grades or similar language ability scores. In the current study it was clear that LD measures which were computed on texts that were lemmatized on the basis of the word family (and thus erased information about test takers' use of derivational affixes) were less powerful in predicting CEFR levels as well as the different Pearson scores (an overall score, a writing score and a vocabulary score) than those that were computed on the basis of the lemma (which only erased inflections). It is possible, however, that studies which are done for different purposes

require different operationalisations of the basic concept of types. In Treffers-Daller and Korybski (2016) we needed to make LD measures comparable across typologically very different languages (Polish and English). In such cases, using the word family as the unit of analysis, and lemmatizing on that basis is probably a good idea. Clearly the typological characteristics of a language will need to be considered carefully before decisions regarding the most appropriate lemmatization strategy can be taken.

Counting *work* and *worker* as different types may be particularly important if one works with data from non-native speakers, as was the case in the current project, because the links between the roots and the derived forms are less strong in L2 than in L1 (Silva and Clahsen 2008). The latter argue that L2 processing relies less on morphological decomposition than L1 processing. If this is indeed the case, there is no justification for erasing the derivational affixes in L2 writing, because knowledge of the root does not automatically entail knowledge of the derived forms (Schmitt and Meara 1997; Schmitt and Zimmermann 2002).

The study also provided important information about the ability of different measures of LD to discriminate between different levels of the CEFR. The basic measures were found to explain more variance in CEFR levels and were more powerful predictors of the Pearson scores. It was particularly revealing that the most basic of all measures, a simple count of lemmatized types, where inflected forms of words are counted as tokens of the same type but derived forms are considered as separate types, turned out to explain most of the variance. Thus, transforming a basic count of types into a TTR or an Index of Guiraud was not helpful. Guiraud is a mathematical transformation of the TTR with a root function but it is unclear whether or not Guiraud overcompensates or undercompensates for the systematically falling TTR. With increasing text length the Guiraud curve flattens out (Daller 2010). Because of the strict compensation formula, Guiraud might not able to account for subtle differences

between learners. Despite its debatable theoretical basis, Guiraud seems to be the most stable measure for language learner data among a series of mathematical transformations of the TTR (van Hout and Vermeer 1988). The much more complex formulae on which the sophisticated measures are based were even less useful. In addition, disambiguating between homographs (e.g. infinitival *to* and prepositional *to*) turned out not to be helpful in discriminating between CEFR levels. In summary, the current project provides clear evidence for Occam's razor, namely that one should not multiply entities unnecessarily, or put differently, the simplest explanation for some phenomenon is more likely to be accurate than more complicated explanations. Our study confirms the results of Grobe (1981) who found that the number of types explained 29.3% of variance in holistic ratings of essays of 750 grade eleven students, which made this the best predictor of the holistic ratings among 23 predictor variables. Richards and Chambers (1996) also found that a simple measure of lexical range (number of different words) correlated significantly with teacher judgements of lexical range. If with our automated measures we want to emulate teachers' ratings, it is important that we stay as close as possible to what raters actually do. It may well be, as Tonkyn (2012) suggests, that raters (sub)consciously count the number of different words in a text[19]. If so, a measure of LD which is based on such type counts has strong ecological validity. The fact that L1 and L2 users are aware of the frequency of words (Schmitt and Dunham 1999) lends support to the assumption that raters count the number of different words in an essay and could thus make use of this information in judging essay quality. It would be much less plausible to claim that raters use complex formulae such as those needed to compute D or HD-D in assessing essays. Support for the ecological validity of this basic measure of LD can also be obtained from the longitudinal analyses of Agatha Christie's

[19] As one reviewer points out, judges may also pay attention to the complexity of words (e.g. derivations) in rating the vocabulary in a text.

novels (Lancashire and Hirst 2009) which revealed that a drop in the number of different words over time in the written work of an individual is a powerful diagnostic tool for dementia. It cannot be emphasised enough, however, that simple counts of types can only be used to measure LD if text length is kept constant. It is not possible to compare number of types across texts of different lengths, although this still happens in some studies (Normand, Parisse, and Cohen 2008).

Many studies in the field of LD over the past 100 years have focused on finding a measure that is not text-length dependent. Such a Holy Grail has not yet been found, however. McCarthy and Jarvis (2007) suggest to identify stable ranges within which it is possible to compare scores of LD. For D, for example, scores can be reliably compared across a stable range of 100 and 400 tokens, whilst MAAS should be relatively stable from 200 to 666 words (McCarthy & Jarvis 2010). Although in the current study the range in text length was very small by comparison with the limits given above for MAAS and D, namely 187-357, this would not have been a safe range within which to compute LD measures. In fact, in the current project, text length was a major predictor of the Pearson scores in regression analysis. The safest option therefore remains to keep text length constant. This is, in fact, the most common solution in research designs where researchers want to prevent a particular variable from exerting influence on the dependent variable (Cohen, Manion, and Morrison 2013, 66)

5. Conclusion

The current study has shown that measures of LD are very useful tools in automated analyses of students' vocabulary in essays, but it is also clear that on their own they cannot distinguish between the levels of the CEFR. However, it would be unrealistic to expect a single measure to discriminate between these levels, because the CEFR levels overlap to a

large extent (see Figure 1). This does not invalidate the use of LD measures which are widely used as a general purpose measure of spoken and written language development (Malvern et al., 2004, p. 8).

Further research will need to focus on which operationalisation of types is most suitable for different purposes, and for different authors/speakers. The current study has provided some evidence for the claim that for bilinguals/L2-users a lemma-based operationalisation is better. Support for this position comes from studies which show that bilinguals or L2-users do not necessarily know the derived forms associated with particular roots. For monolinguals a different operationalisation of types may be suitable, because there are stronger links between the roots and the derived forms for monolinguals than for bilinguals, according to Silva and Clahsen (2008). The results regarding the different lemmatization principles are probably the most relevant finding for language testers interested in identifying further criterial features for the different levels of the CEFR. The current study mainly focused on derivational affixes up to level 3 in Bauer and Nation (1993) but it is clear that derivational affixes beyond this level are also being used by L2 learners, for example the *–th* suffix as in *growth* or *strength*. To what extent other affixes from levels 4-6 can help discriminate between CEFR levels is an interesting point worth pursuing in future research. Finally, future research in this field should consider analysing formulaic language, that is fixed phrases at different levels of complexity, because formulaic language has been shown to be fundamental to the ways in which language is used and processed (Martinez and Schmitt 2012). An obvious limitation of LD measures is that they are based on analyses of single words and do not take into account formulaic language.

For researchers and language teachers interested in evaluating the difficulty of texts it is important to be aware of the risks of computing measures of lexical diversity without keeping text length constant. In addition, it may be of interest to know that TTRs or D values

computed by different pieces of software will differ depending on the programme's operationalisation of the notion "different word" (type). Vocabprofile (http://www.lextutor.ca/vp/comp/), for example, which is widely used by researchers and teachers, computes TTRs without lemmatization, which means that inflected forms of verbs (as in *walks* and *walked*) and singular and plural forms of nouns (*car* and *cars*) are considered as different words (types). While there may be good arguments for doing this, computations which involve lemmatizations based on the lemma or on the word family as the unit of analysis would produce lower scores.

To help researchers interpret and evaluate the scores obtained with the help of LD measures, we urgently need LD norms for different text genres, for spoken and written language, and for different speaker groups (monolinguals, bilinguals and L2 learners at ages or different proficiency levels). As Olinghouse and Wilson (2013) have shown, narrative texts are more diverse than informative or persuasive texts, and L2 learners have also been found to be more lexically diverse in dialogues than in monologues (Michel 2011). In the current study test takers from a wide variety of countries and L1 backgrounds took part. It is possible that the L1 of bilinguals also affects the LD scores. As one reviewer points out, learners whose L1 is similar to the L2 might obtain higher LD scores than those for whom this is not the case. In the current study we were not able to pursue this matter but this would need to be taken into consideration in future studies aiming to develop LD norms for different groups of speakers and writers. Finally, more studies of LD in non-Indo-European languages are needed to ensure the field benefits from insights from a wide variety of typologically different languages.

References

Alderson, C. 2007. 'The CEFR and the need for more research,' *The Modern Language Journal* 91: 659-663

Bartning, I., M. Martin, and I. Vedder (eds.). 2010. *Communicative proficiency and linguistic development: Intersections between SLA and language testing*. Rome: Eurosla.

Bulté, B., and A. Housen. 2012. 'Defining and operationalising L2 complexity' in A. Housen, F. Kuiken and I. Vedder (eds.) *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. John Benjamins, pp. 21-46.

Cohen, L., L. Manion, and K. Morrison. 2013. *Research methods in education*. Routledge.

Crossley, S.A., and D.S. Macnamara. 2011. 'Shared features of L2 writing: Intergroup homogeneity and text classification,' *Journal of Second Language Writing* 20: 271-285.

Crossley, S.A., T. Salsbury, and D.S. Macnamara. 2009. 'Measuring second language lexical growth using hypernymic relationships,' *Language Learning* 59/2: 307-334.

Crossley, S.A., T. Salsbury, and D.S. Macnamara. 2013. 'Validating lexical measures using human scores of lexical proficiency,' in S. Jarvis and M. H. Daller (eds.) *Human ratings and automated measures*. John Benjamins, pp. 105-134.

Crossley, S.A., T. Salsbury, and D.S. Macnamara. 2014. 'Assessing lexical proficiency using analytic ratings: a case for collocation accuracy.' *Applied Linguistics*. Online first:1-22.

Daller, M. 2010. 'Guirauds index of lexical richness.' Oral presentation at the Annual Meeting of the British Association of Applied Linguistics, September 2010.

Daller, H., R. van Hout, R. and J. Treffers-Daller. 2003. 'Measuring lexical aspects of oral language proficiency among bilinguals: an analysis of different measurements', *Applied Linguistics* 24/2: 197-222.

Durán, P., D. Malvern, J.B. Richards, and N. Chipere. 2004. 'Developmental trends in lexical diversity.' *Applied Linguistics* 25/2: 220-242.

Engber, C.A. 1995. 'The relationship of lexical proficiency to the quality of ESL compositions.' *Journal of Second Language Writing* 4/2:139-155.

Grobe, C. 1981. 'Syntactic maturity, mechanics, and vocabulary as predictors of quality ratings.' *Research in the Teaching of English* 15/1:75-85.

Guiraud, P. 1954. *Les caractéristiques statistiques du vocabulaire*. Presses Universitaires de France.

Housen, A., F. Kuiken, and I. Vedder (eds.). 2012. *Dimensions of L2 performance and proficiency: complexity, accuracy and fluency in SLA*. John Benjamins.

Hulstijn, J. 2007. 'The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency,' *The Modern Language Journal* 91: 663-667.

Hulstijn, J., R. Schoonen, N. H De Jong, M.P. Steinel, and A.F. Florijn. 2012. 'Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR)." *Language Testing* 29: 202-220.

Jarvis, S. 2002. 'Short texts, best fitting curves, and new measures of lexical diversity.' *Language Testing* 19: 57-84.

Jarvis, S. 2013. 'Defining and measuring lexical diversity.' in S. Jarvis and M.H. Daller (eds.) *Vocabulary knowledge: human ratings and automated measures*. John Benjamins, pp. 13-44.

Johnson, W. 1944. 'Studies in language behavior 1: a program of research,' *Psychological Monographs* 56: 1-15.

Kobrin, J.L., H. Deng, and E.J. Shaw. 2011. 'The association between SAT prompt characteristics, response features, and essay scores.' *Assessing Writing* 16: 154-169.

Lancashire, I. and G. Hirst. 2009. 'Vocabulary changes in Agatha Christie's mysteries as an indication of dementia: A case study'. Paper presented at the 19th Annual Rotman Research Institute Conference, Cognitive Aging: Research and Practice, 8–10 March 2009, Toronto.

Laufer, B. 1991. 'The development of L2 lexis in the expression of the advanced language learner.' *Modern Language Journal* 75/4: 440-448.

Laufer, B., and I.S.P Nation. 1995. 'Vocabulary size and use: lexical richness in L2 written production.' *Applied Linguistics* 16: 307-322.

Maas, H.D. 1972. 'Zusammenhang zwischen Wortschatzumfang und Länge eines Textes.' *Zeitschrift für Literaturwissenschaft und Linguistik* 8: 73-79.

MacWhinney, B. 2000. 'The CHILDES project: Tools for analyzing talk: Volume I: Transcription format and programs, volume II: The database.' *Computational Linguistics* 26/4: 657-657.

Malvern, D., J.B. Richards, N. Chipere, and P. Durán. 2004. *Lexical richness and language development: Quantification and assessment*. Palgrave MacMillan.

Martinez, R. and N. Schmitt. 2012. A phrasal expressions list. *Applied Linguistics* 2012: 33/3: 299–320.

McCarthy, P.M. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. University of Memphis, Memphis, USA.

McCarthy, P.M., and S. Jarvis. 2007. 'A theoretical and empirical evaluation of vocd,' *Language Testing* 24: 459–488.

McCarthy, P.M., and S. Jarvis. 2010. 'MTLD, Vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment.' *Behavior Research Methods* 42: 381-392.

Michel, M.C. 2011. 'Effects of task complexity and interaction on L2 performance.' In P.Robinson (ed.). *Second language task complexity: Researching the cognition hypothesis of language learning and performance*. John Benjamins, pp. 141-174.

Nation, I.S.P, and L. Bauer. 1993. 'Word families.' *International Journal of Lexicography* 6/4: 253-279.

Normand, M.T., C. Parisse, and H. Cohen. 2008. 'Lexical diversity and productivity in French preschooolers: developmental, gender and sociocultural factors.' *Clinical Linguistics and Phonetics* 22/1: 47-58.

Olinghouse, N.G., and J. Wilson. 2013. 'The relationship between vocabulary and writing quality in three genres.' *Reading and Writing* 26: 45-65.

Richards, J.B., and F. Chambers. 1996. 'Reliability and validity in the GCSE oral examination.' *Language Learning Journal* 14: 28-34.

Schmitt, N., and B. Dunham. 1999. 'Exploring native and non-native intuitions of word frequency.' *Second Language Research* 15/4: 389-411.

Schmitt, N. and P. Meara. 1997. 'Researching vocabulary through a word knowledge framework. Word associations and verbal suffixes.' *Studies in Second Language Acquisition* 20: 17-36.

Schmitt, N., and C. B. Zimmerman. 2002. 'Derivative word forms: what do learners know?' *TESOL Quarterly* 36/2:145-171.

Silva, R., and H. Clahsen. 2008. 'Morphologically complex words in L1 and L2 processing: evidence from masked priming experiments in English.' *Bilingualism: Language and Cognition* 11: 245-260.

Staehr, L.S. 2008. Vocabulary size and the skills of listening, reading and writing,' *Language Learning Journal* 36/2: 139-152.

Templin, M. 1957. *Certain language skills in children*. Minneapolis: University of Minneapolis.

Thomson, G.H., and J.R. Thompson. 1915. 'Outlines of a method of the quantitative analysis of writing vocabularies.' *British Journal of Psychology* 8: 52-69.

Tonkyn, A.P. 2012. 'Measuring and perceiving changes in oral complexity, accuracy and fluency,' in A. Housen, F. Kuiken and I. Vedder (eds.) *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*. John Benjamins, pp. 221–244.

Treffers-Daller, J. 2013. 'Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability.' In: S. Jarvis. S. and M. Daller, M. (eds.) *Vocabulary Knowledge: Human Ratings and Automated Measures*. John Benjamins, pp. 79-105.

Treffers-Daller, J. and T. Korybski. 2016. 'Using lexical diversity measures to operationalise language dominance in bilinguals.' in C. Silva-Corvalán, and J. Treffers-Daller (eds.) *Language dominance in bilinguals: issues of operationalization and measurement*. Cambridge University Press, pp. 106-133.

Van Hout, R. and A. Vermeer. (1988). Spontane taaldata en het meten van lexicale rijkdom in tweede-taalverwerving. *Toegepaste Taalwetenshap in Artikelen* 32: 108 – 122.

Vermeer, A. 2000. 'Coming to grips with lexical richness in spontaneous speech data.' *Language Testing* 17/1: 65-83.

Wu, T. 1993. 'An accurate computation of the hypergeometric distribution function.' *ACM Transactions on Mathematical Software* 19: 33-43.

Yu, G. 2010. 'Lexical diversity in writing and speaking task performances.' *Applied Linguistics* no. 31/2: 236-259.

Table 1. Students' level of competence according to the CEFR

CEFR level	**B1**	**B2**	**C1**	**C2**
N	50	50	50	29

Table 2a. Mean scores on basic measures of LD across different levels of the CEFR

Measures	B1	B2	C1	C2	Overall means and SD	Eta Squared
Types 0	101.52	109.48	111.66	114.76	108.72 (9.98)	.225
Types 1	96.32	104.14	106.32	109.48	103.43 (9.82)	.229
Types 2	96.24	103.92	106.06	109.07	103.21 (9.87)	.221
TTR 0	0.56	0.61	0.61	0.63	0.60 (0.06)	.229
TTR 1	0.52	0.57	0.58	0.60	0.60 (0.06)	.248
TTR2	0.52	0.57	0.58	0.59	0.56 (0.06)	.234
Guiraud 0	7.51	8.14	8.27	8.48	8.05 (0.74)	.232
Guiraud 1	7.09	7.71	7.86	8.08	8.03 (0.74)	.242
Guiraud 2	7.08	7.69	7.84	8.04	7.50 (0.73)	.230

Table 2b. Mean scores on sophisticated measures of LD across different levels of the CEFR

measures	B1	B2	C1	C2	Overall means and SD	Eta Squared
D (VOCD) 0	72.40	85.71	86.61	89.54	82.86 (21.29)	.092
D (VOCD) 1	61.88	71.65	73.83	76.61	70.33 (17.28)	.098
D (VOCD) 2	62.20	71.58	74.48	75.67	70.15 (17.25)	.085
HDD 0	34.47	35.37	35.51	35.64	35.21 (1.40)	.109
HD-D 1	33.55	34.29	34.55	34.75	34.23 (1.39)	.100
HD-D2	33.61	34.36	34.36	34.86	34.30 (1.40)	.104
MTLD 0	70.14	84.55	88.47	93.85	83.12 (22.96)	.134
MTLD 1	58.70	68.52	72.81	77.11	68.37 (17.06)	.140
MTLD 2	59.68	70.01	73.69	78.92	69.60 (17.82)	.145

*0 = no lemmatization, 1 = first lemmatization principle, 2 = second lemmatization principle

Table 3. ANOVA and Tukey post hoc test results for LD measures (first lemmatization principle) across different levels of the CEFR

	F	p	B1-B2	B1-C1	B1-C2	B2-C1	B2-C2	C1-C2
Types	17.034	<.0001	*	*	*	ns	ns	ns
TTR	18.923	<.0001	*	*	*	ns	ns	ns
Guiraud	18.270	<.0001	*	*	*	ns	ns	ns
D (VOCD)	6.198	.0005	ns	ns	*	ns	ns	ns
HD-D	6.388	.0004	ns	ns	*	ns	ns	ns
MTLD	9.757	<.0001	ns	*	*	ns	ns	ns

For post hoc comparisons, alpha was set at .0014[20]

Table 4. Correlations between LD measures and Pearson scores

	TTR	Guiraud	D	HD-D	MTLD	Vocab score	Writing score	Overall score
types	.973**	.993**	.840**	.843**	.783**	.468**	.447**	.470**
TTR		.993**	.857**	.860**	.787**	.470**	.424**	.455**
Guiraud			.854**	.858**	.790**	.472**	.438**	.466**
D				.925**	.794**	.319**	.290**	.314**
HD-D					.827**	.309**	.276**	.299**
MTLD						.331**	.344**	.338**
Vocab							.765**	.804**
Writing								.920**

[20] Because of the larger number of comparisons (6 variables across 4 levels of the CEFR, which resulted in 6 * 6 post hoc tests = 36 comparisons), alpha was set at .05/54 = 0.0014.

Figure 1 Lexical complexity at different levels of construct specification (from Bulté & Housen, 2012, p. 28)

Permission to reproduce the Figure has been granted by John Benjamins Publishing Company

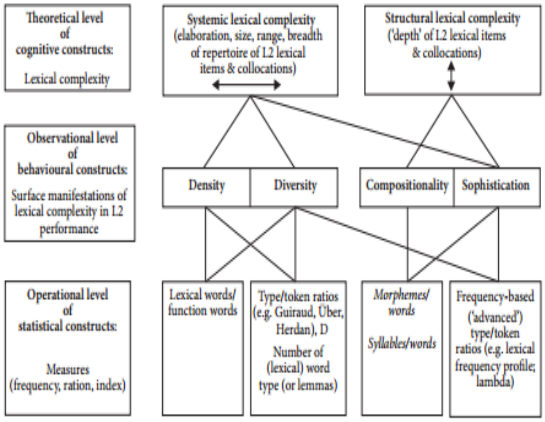

Figure 2. Guiraud (Lemmatization 1) values across different levels of the CEFR